Modélisation explicite des objets et de l'environnement en combinant les approches topologique et métrique pour la localisation

M. Decrouez[1,2] R. Dupont[1] F. Gaspard[1] F. Devernay[2] J.L. Crowley[2]

[1] CEA, LIST, Laboratoire Vision et Ingénierie des Contenus
[2] INRIA Grenoble - Rhône-Alpes

[1] Point Courrier 94, Gif-sur-Yvette, F-91191 France
[2] Avenue de l'Europe, 38330 Montbonnot-Saint-Martin
marion.decrouez@cea.fr

Résumé

Nous présentons dans cet article une nouvelle formulation de la modélisation et de la reconnaissance de lieu. Les milieux intérieurs sont composés d'une quantité d'objets susceptibles d'être déplacés. Nous souhaitons exploiter les multiples passages d'une caméra dans un même environnement pour modéliser d'une part la structure de la scène et d'autre part les objets le constituant. Nous proposons une association de méthodes de SLAM métrique et topologique pour détecter et représenter les objets, ainsi qu'un formalisme probabiliste pour la reconnaissance de lieu prenant en compte l'évolution des objets.

Mots Clef

SLAM par vision, sac de mots, reconnaissance de lieu, objets.

Abstract

This paper presents a new formulation of place recognition in unknown environments. Many objects in indoor environments are likely to be moved. We want to make the most of several observations of a camera in the same scene to represent the different places and objects. We propose to combine methods of metrical and topological localization to detect and model objects. We present also a probabilistic approach for place recognition that considers the objects.

Keywords

Visual SLAM, bag of words, place recognition, objects.

1 Introduction

De récents travaux en localisation et cartographie simultanées utilisant uniquement des capteurs visuels (*visual SLAM : visual simultaneous localization and mapping*) permettent à un système monoculaire de créer en temps réel un modèle de l'environnement sans aucune connaissance a priori. La position de la caméra est estimée à chaque instant dans l'environnement reconstruit. Les systèmes reposant sur l'apprentissage des paramètres géométriques de la scène (slam métrique) [15], [13] offrent de nombreuses possibilités d'applications en réalité augmentée. Les travaux actuels permettent d'incruster de façon réaliste des objets virtuels dans la séquence d'images et intéressent l'industrie des jeux videos. L'aprofondissement de ces systèmes permettront à l'avenir des applications industrielles précises comme l'aide à la maintenance ou à l'assemblage. Ces méthodes ont l'inconvénient de ne pas fournir d'informations sémantiques sur l'environnement observé : nous ne reconnaissons ni le lieu ni les objets. Pourtant, dans le cadre de scénarios mettant en scène une intéraction de l'utilisateur avec son milieu, il est primordial de comprendre le contexte dans lequel évolue la caméra et d'identifier les objets présents dans l'environnement. Des approches de SLAM reposant uniquement sur l'apparence ou SLAM topologique [25], [20], [7], [2] classifient les images en lieux que l'on peut catégoriser [24] (milieu extérieur ou intérieur, cuisine, couloir, bureau, rue). Ces méthodes s'affranchissent des problèmes d'accumulation d'erreurs des systèmes de SLAM métrique et déterminent dans quel lieu ou quelle pièce se trouve l'utilisateur. Elles répondent efficacement au problème de détection de fermeture de boucle. Néanmoins elles ne permettent pas d'établir une pose précise de la caméra et des points de repère de la scène. Elles sont donc inadaptées aux applications de réalité augmentée citées précédemment. La reconnaissance de lieu est sujette à d'autres problèmes. L'existence de structures répétitives, ambigues et non discriminantes dégrade la qualité des résultats. Elle est aussi perturbée par la présence d'objets susceptibles d'être déplacés, notamment en milieu intérieur. La modélisation de milieux dynamiques comme les rayons d'un supermarché ou les ateliers d'une usine doit contrôler ces perturbations. Elle peut faire intervenir la collaboration de plusieurs utilisateurs. Dans ce cas, le traitement des informations provenant de différentes sources nécessite de classifier l'environnement en lieux pour mieux détecter les différences et décrire l'évolution du milieu.

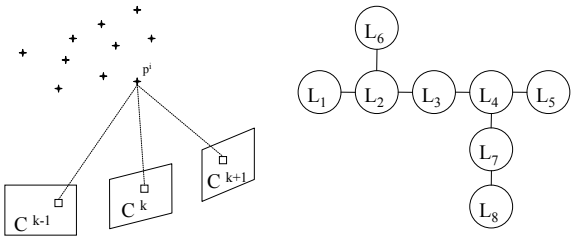
L'approche présentée dans cet article propose de définir explicitement la scène comme une structure statique et un ensemble d'objets dynamiques. Nous proposons un nouveau modèle qui enregistre les coordonnées des points statiques de l'environnement et représente les points dynamiques comme des objets. Sans autre information a priori, un objet est définit comme un ensemble de primitives visuelles ayant eu le même déplacement par rapport à la structure statique. Notre objectif est d'exploiter les multiples passages de la caméra dans un même environnement et de tirer le maximum d'informations sur la scène et son évolution au fil du temps. Les algorithmes de reconnaissance d'images permettent de détecter un lieu déjà visité. La reconstruction 3D de l'environnement met en évidence le mouvement de certains points : leur géométrie ne correspond pas à celle de la structure statique. Ces *outliers* sont des informations utiles que nous souhaitons explicitement utiliser. En associant de cette façon les deux approches nous inférons la présence d'objets dont nous pouvons apprendre le modèle. La section 2 donne une vue d'ensemble des approches de SLAM métrique et topologique. La section 3 détaille la représentation des images et des lieux de la carte topologique. La section 4 décrit le filtrage bayésien utilisé pour la reconnaissance de lieu. La section 5 montre la possibilité d'apprendre des modèles d'objets et nos premiers résultats.

2 Etat de l'art

Il existe aujourd'hui de nombreuses approches proposant des solutions au problème du SLAM. Les approches fondées uniquement sur la vision (Vision-SLAM) peuvent être classées en deux catégories : les approches métriques et les approches topologiques. Les algorithmes de SLAM métrique (fig. 1(a)) permettent de reconstruire un environnement inconnu et de s'y localiser précisément. La carte enregistre les positions géométriques des points de l'environnement et la trajectoire suivie par la caméra. Les méthodes de SLAM topologiques et de reconnaissance de lieux (fig. 1(b)) reposent sur une représentation discrète de l'environnement modélisé sous forme d'un graphe. Les nœuds du graphe sont des lieux distincts et les arêtes représentent les relations entre lieux (positionnement relatif, adjacence temporelle).

2.1 SLAM métrique

Plusieurs travaux ont montré des résultats remarquables dans l'estimation des paramètres d'une scène 3D. Les méthodes fondées sur l'utilisation d'un filtre de Kalman étendu [9] d'une part, et l'optimisation des paramètres par ajustement de faisceaux [15] d'autre part, ont prouvé leur efficacité ces dernières années. Strasdat et al. [23] montrent que les systèmes temps réel s'appuyant sur l'utilisation d'images clefs et l'optimisation par ajustement de faisceaux sont plus performants en terme de précision. L'ajustement de faisceaux global est la méthode de reconstruction 3D la plus précise. Elle permet d'optimiser les paramètres des points reconstruits et les poses des caméras



(a) SLAM métrique (b) SLAM topologique

FIGURE 1 – **(a)** L'algorithme de SLAM métrique calcule les coordonnées des points p en 3D et les paramètres de pose des caméras C. **(b)** En utilisant une carte topologique, l'environnement est décrit par un ensemble de lieux discrets L.

en minimisant l'erreur de reprojection globale. Cette méthode met en œuvre l'optimisation de l'ensemble des paramètres de la scène. Elle est couteuse et n'est pas adaptée aux longues séquences et au traitement en temps réel. Royer et al. [19] proposent de procéder de manière hiérarchique en découpant la séquence en sous-séquences. Les sous-séquences sont traitées puis fusionnées et l'étape finale optimise la séquence globale. L'approche de Mouragnon et al. [15] permet une reconstruction en temps réel avec une méthode d'ajustement de faisceaux local : seuls les paramètres des dernières caméras sont optimisés. Klein et Murray [13] atteignent le temps réel avec PTAM (Parallel Tracking and Mapping). Ils séparent le suivi de la caméra et la cartographie de l'environnement en deux processus parallèles, la cartographie utilise aussi une méthode d'ajustement de faisceau local. Malgré ces progrès récents, il existe encore des limitations aux méthodes de SLAM métrique. On observe une dérive cumulative dans l'estimation de la position de la caméra. Les méthodes de SLAM monoculaire présentent une dérive du facteur d'échelle. Dans ces conditions il est difficile de détecter qu'une zone a déjà été visitée par la caméra. Ces détections de boucle pourraient pourtant améliorer les estimations des positions calculées et de la carte en permettant une correction des erreurs accumulées. De même, si la position de la caméra est perdue, les algorithmes cités ne permettent pas de se relocaliser par rapport aux lieux déjà visités.

2.2 SLAM topologique

Motivées par la détection de fermeture de boucle, de nouvelles approches ont été proposées dans la littérature. Elles souhaitent s'abstraire des problèmes d'accumulation d'erreurs et répondre aux limitations du SLAM métrique. Elles reposent sur une représentation topologique de l'environnement qui découpe l'espace en lieux discrets et distincts. Il ne s'agit pas de se positionner précisément par rapport aux points de l'environnement mais de caractériser l'apparence d'un lieu de manière générale.
Ces méthodes considèrent le problème du SLAM comme un problème de reconnaissance d'images : deux images si-

milaires proviennent probablement du même endroit. En 2006, Ho et Newman [11] proposent une méthode de détection de fermeture de boucle : le système encode la similarité de chaque paire d'images d'une séquence vidéo dans une matrice de similarité, un traitement ultérieur extrait les séquences significatives d'images similaires. La décomposition en valeurs singulières de la matrice élimine les effets des régions ambigues. Les vecteurs singuliers correspondent à des structures dominantes de l'environnement (végétation, façade de batîment, rue). La soustraction de ces modes diminue les effets de ces structures et rend la détection de boucle plus fiable. Les auteurs répondent ainsi au problème d'aliasing perceptuel (deux lieux distincts peuvent avoir la même apparence) mais l'algorithme est lent : le système proposé fournit une localisation globale hors ligne. De nombreuses approches proposent un système de localisation fondée sur une représentation de l'image en sac de mots visuels [25], [11], [7], [2]. Ces méthodes inspirées des techniques de recherche d'information présentent l'image comme un ensemble de primitives visuelles, les mots, définis dans un dictionnaire ou vocabulaire (Sivic et Zisserman 2003 [22], Nister et Stewenius 2006 [16]). En général le dictionnaire est appris hors ligne : des primitives visuelles sont extraites d'un ensemble d'images d'entrainement et regroupées suivant leur similarité pour former les mots visuels. L'image devient un histogramme de mots visuels et la similarité entre deux images est mesurée par la distance entre leurs histogrammes respectifs. Les travaux de Cummins et Newman [7], [8] définissent un formalisme probabiliste reposant sur l'approche en sac de mots visuels (figure 2(a)). L'environnement est un ensemble de lieux discrets, dont l'apparence est modélisée par une distribution sur les mots du dictionnaire. Á chaque instant la nouvelle image est convertie en sac de mots. Pour chaque lieu déjà enregistré dans la carte, la méthode détermine la probabilité que l'image courante corresponde au modèle d'apparence du lieu. Le système estime la provenance de l'image courante au sens du maximum *a posteriori* et le résultat obtenu permet de se localiser et de mettre à jour le modèle, ou d'apprendre le modèle d'un nouveau lieu. L'algorithme offre une robustesse remarquable à l'aliasing perceptuel grâce à la prise en compte des probabilités de co-occurrences des mots visuels (calculées hors-ligne) dans l'estimation de la vraisemblance de l'observation. Les temps de calcul sont cependant couteux et n'autorisent pas de traitements en temps réel. Les travaux d'Angeli et al. [2] permettent de détecter en temps réel qu'une image provient d'un lieu déjà visité. Leur approche repose aussi sur une représentation de l'image en sac de mots (figure 2(b)) et la probabilité de fermeture de boucle est estimée par filtrage bayésien. Contrairement aux méthodes précitées, le vocabulaire utilisé est appris en ligne de façon incrémentale à partir d'une structure initialement vide. Leur système est robuste à l'aliasing perceptuel grâce à une vérification de la consistence géométrique. Notre algorithme de reconnaissance de lieu s'appuie en partie sur ces travaux.

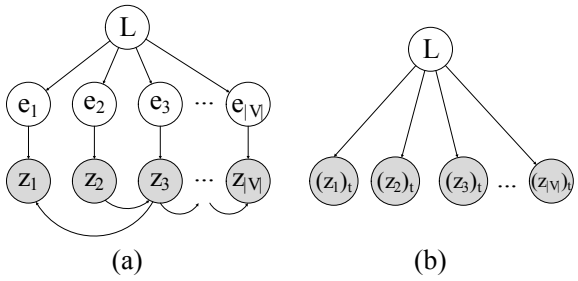


FIGURE 2 – La variable L désigne un lieu. e_i modèle l'existence du mot i, z_i modèle l'observation du mot i. **(a)** Modèle graphique du système proposé par Cummins et Newman [7] : un lieu est modélisé par le modèle d'apparence donné par $\{p(e_i = 1|L)\}_{i=1}^{|V|}$. La détection z_i est reliée à l'existence e_i en prenant en compte la qualité du détecteur de mots visuels. Les observations z_i sont interdépendantes. **(b)** Modèle graphique des travaux d'Angeli et al. [2] : un lieu est modélisé par une distribution sur les mots visuels d'un dictionnaire dépendant du temps t.

2.3 Localisation dans un milieu connu

Nous souhaitons proposer un outil de collaboration permettant de modéliser un environnement. Nous devons pour cela nous localiser par rapport aux données apprises lors de précédents passages de la caméra. Des algorithmes de localisation en milieu connu sont utilisés lorsque des *a priori* sur l'environnement sont disponibles. Schindler et al. [21] proposent de se relocaliser dans une base de données géoréférencées contenant 30000 images d'une ville. La performance du système s'explique par l'utilisation d'un arbre de vocabulaire pour la mise en correspondance des primitives visuelles. Arnold et al. [12] et Gay-Bellile et al. [10] utilisent un nuage de points 3D géoréférencés (construit lors d'une phase d'apprentissage) pour se relocaliser ou réduire la dérive due aux erreurs accumulées. Nous réutilisons l'idée de se localiser dans un nuage de points 3D pour détecter des points ne vérifiant pas la consistance géométrique de la scène.

2.4 Méthodes utilisant la reconnaissance d'objet

Nous proposons un système utilisant la reconnaissance de lieu et d'objets pour modéliser notre environnement. Plusieurs méthodes de SLAM utilisent déjà la reconnaissance d'objets. Elles identifient en général des objets d'une base de données apprise au préalable. Ahn et al. [1] construisent une base de données d'objets plans dans un processus hors ligne et se servent de ces objets comme points de repères pour une localisation utilisant un filtre de Kalman étendu. L'algorithme est ainsi plus robuste et permet une implémentation temps réel en réduisant le nombre de points reconstruits dans l'environnement. Seuls les points des objets reconnus sont reconstruits. Castle et al. [6] reconstruisent

l'environnement suivant la méthode MonoSLAM de Davison et al. [9], détectent la présence d'objets connus et les localisent. Ces deux approches nécessitent d'apprendre préalablement un modèle des objets : l'ensemble des descripteurs SIFT de chaque objet est chargé dans une base de données. L'algorithme de reconnaissance (mise en correspondance des descripteurs et vérification de la géométrie) permet d'identifier seulement quelques dizaines d'objets différents. Cette méthode n'est pas adaptée à des base de données de grande taille. Reitmayr [18] propose d'enrichir la base en ligne. L'utilisateur explore l'environnement et segmente manuellement l'objet qu'il souhaite reconnaître lors d'un deuxième passage. Les auteurs de [5] propose de corriger l'estimation du facteur d'échelle (SLAM métrique) avec la reconnaissance d'objets de taille connue. Enfin les auteurs de [3] proposent de regrouper les primitives observées suivant leur similarité et leur proximité en *clusters* 3D . Ces *clusters* sont souvent associés à des objets réels de la scène et pourraient être utilisés pour une reconnaissance de lieu robuste.
Nous proposons une nouvelle formulation de la reconnaissance de lieu. Notre approche s'appuie sur la représentation en sac de mots visuels présentée dans la section suivante.

3 Représentation de l'apparence

La représentation d'une image en sac de mots s'inspire des travaux de Sivic et Zisserman 2003 [22] et Nister et Stewenius 2006 [16].

3.1 Détection de points d'intérêt et description

L'image est représentée comme un ensemble de primitives visuelles locales. Ce mode de représentation offre une description de l'image robuste aux occultations partielles et adaptée aux applications suivantes. Le suivi des primitives locales dans plusieurs images consécutives permet de calculer les positions géométriques des points détectés dans l'environnement et la trajectoire de la caméra. Les primitives d'une image peuvent aussi être quantifiées en mot visuels et fournir une description qualitative et globale de l'apparence. Nous utilisons le détecteur de Harris-Stephen et le descripteur SURF proposé par Bay et Tuytelaars dans [4] en raison de leur efficacité et performance.

3.2 Sac de mots visuels

L'environnement est décrit par un ensemble de lieux discrets. Nous souhaitons utiliser un modèle du lieu retenant les informations pertinentes et permettant une mise à jour. Le lieu est caractérisé initialement par la description d'une image et le modèle doit intégrer par la suite les informations provenant d'images similaires. Nous utilisons donc une représentation en sacs de mots visuels, méthode très utilisée en catégorisation et reconnaissance d'images qui permet une mise à jour du modèle du lieu.

Description d'une image. Une image ou une observation est un histogramme d'occurrences des mots visuels du dictionnaire : les descripteurs extraits de l'image sont quantifiés en mots visuels suivant une quantification dite *douce*, inspirée de [17]. On détermine pour chaque descripteur extrait les trois plus proches voisins dans le dictionnaire. Les mots trouvés contribuent à l'histogramme suivant leur distance au descripteur. On note x le descripteur de l'image et $\{c_k\}_{k=1}^{3}$l'ensemble des cinq mots du dictionnaire retenus, le poids attribué à chaque mot vaut :

$$w_i = \frac{\exp\left(-d\left(x, c_i\right)\right)}{\sum_{k=1}^{3} \exp\left(-d\left(x, c_k\right)\right)} \quad (1)$$

$d(x, c_i)$ est la distance euclidienne entre le descripteur x et le mot du dictionnaire c_i. C'est la distance utilisée par les auteurs de [4] pour la mise en correspondance des points d'intérêt. Ce mode de quantification permet de garder un maximum d'information tout en minimisant les erreurs d'assignement. En effet, en assignant chaque descripteur à un seul mot, la présence de structures répétitives produit un aliasing perceptuel (des lieux distincts ont la même description). Une solution est de garder les appariemments vérifiant le critère de Lowe [14] (ratio des distances au deux plus proches voisins inférieurs à un seuil) mais on perd ainsi de l'information. En utilisant cette approche, la quantification est plus précise et la mise en correspondance des images est moins perturbée par la présence de structures répétitives ou ambigues.

Description d'un lieu. Un lieu est associé à un modèle d'apparence correspondant à une distribution sur les mots du vocabulaire.

3.3 Construction du dictionnaire

Le dictionnaire est appris hors ligne. Les descripteurs SURF extraits d'un ensemble d'images sont agglomérés suivant la méthode des k-means pour fournir un dictionnaire de 10000 mots. Les images d'entrainement sont des images aléatoires téléchargées sur Flickr (environ 3000 images utilisées). Les mots du dictionnaire sont ensuite indexés dans une structure d'arbre k-means.

4 Formalisme bayésien pour la reconnaissance du lieu

Nous utilisons un formalisme inspiré de [2]. Le vocabulaire de mots visuels n'est cependant pas appris progressivement lors de l'exploration mais dans un processus hors ligne (section 3.3). Cette méthode permet de couvrir l'espace des descripteurs de manière optimale. Nous utilisons un seul type de primitives visuelles pour la description des images. La quantification des descripteurs SURF en mots visuels est différente. Nous souhaitons en effet quantifier tous les descripteurs sans produire un aliasing perceptuel. Comme décrit dans la partie 3.2, chaque descripteur est associé à une moyenne pondérée de mots visuels et est ainsi décrit de manière plus précise.

4.1 Notations

L'observation courante (au temps k) est notée $Z_k = \left\{z_1, ..., z_{|V|}\right\}$. $|V|$ est la taille du vocabulaire et la variable binaire z_i indique la présence ou l'absence du i-ième mot dans l'image. On note $Z^k = \{Z_1, ..., Z_k\}$ l'ensemble des observations jusqu'au temps k. Au temps k la carte de l'environnement contient l'ensemble des lieux $\{L_1, ..., L_n\}$.

4.2 Estimation du lieu au sens du maximum *a posteriori*

Le système doit dans un premier temps juger si l'image courante provient d'un lieu déjà visité dans le passé et définit dans la carte topologique. Il s'agit de calculer les probabilités d'être dans chacun des lieux de la carte connaissant l'ensemble des observations jusqu'au temps k et d'en déduire le lieu L_j dont l'index vérifie :

$$j = argmax_{i=-1,1,...,n} p\left(L_i|Z^k\right) \tag{2}$$

Le résultat $j = -1$ signifie qu'aucun lieu n'a été reconnu et qu'il faut mettre à jour la carte en ajoutant un nouveau lieu. Le calcul du lieu L_{-1} est expliqué plus loin en section 4.4.

Le lieu est estimé au sens du maximum *a posteriori*. En suivant la loi de Bayes, la probabilité *a posteriori* d'être dans le lieu L_i s'écrit :

$$p\left(L_i|Z^k\right) = \frac{p\left(Z_k|L_i\right) p\left(L_i|Z^{k-1}\right)}{p\left(Z_k|Z^{k-1}\right)} \tag{3}$$

$p\left(L_i|Z^{(k-1)}\right)$ est le terme de prédiction : c'est la probabilité a priori d'être dans le lieu L_i. $p\left(Z_k|L_i\right)$ est la vraisemblance de l'observation : connaissant la carte de l'environnement on évalue pour chaque hypothèse de lieu la vraisemblance des mots $Z_k = \left\{z_1, ..., z_{|V|}\right\}$ observés dans l'image courante. $p\left(Z_k|Z^{k-1}\right)$ est le terme de normalisation.

4.3 Vraisemblance de l'observation

La vraisemblance de l'observation permet de mesurer la similarité entre l'observation courante et les lieux déjà visités. Le calcul des scores de similarité avec les lieux enregistrés se fait pendant la quantification des descripteurs de la nouvelle image. Pour un mot trouvé, constituant le modèle d'apparence d'un lieu, le score de similarité est mis à jour en additionnant un terme inspiré de la méthode de pondération TF-IDF (Term Frequency - Inverse Document Frequency) :

$$tf - idf = p_{w_{L_i}} \log\left(\frac{1}{p_w}\right) \tag{4}$$

p_w est la probabilité d'occurrence du mot w. $\frac{1}{p_w}$ est une mesure de l'importance du mot visuel w. Les mots apparaissant peu sont considérés comme plus discriminants et ont un poids plus important. Ce terme, correspondant à la fréquence de document inverse, est calculé lors de la construction du dictionnaire :

$$\frac{1}{p_w} = \frac{N}{N_w} \tag{5}$$

N est le nombre d'images d'entrainement utilisées pour la construction du dictionnaire. N_w est le nombre d'images d'entrainement contenant le mot w. $p_{w_{L_i}}$ est la probabilité de trouver le mot w dans le lieu Li. Si le lieu est décrit par une image, c'est la fréquence du mot w dans l'image. S'il est décrit par plusieurs images c'est la moyenne des fréquences dans chaque image.

4.4 Nouveau Lieu

Le système doit être capable de déterminer si l'observation provient d'un lieu de la carte ou d'un lieu encore inconnu. Pour cela nous apprenons le modèle d'un lieu virtuel (le lieu L_{-1}) qui figure une moyenne des lieux déjà visités. Le modèle est initialisé de sorte que tous les mots visuels existent avec la probabilité d'occurrence calculée lors de la construction du dictionnaire. Il est ensuite mis à jour en retenant les m mots les plus fréquemment observés (m est le nombre moyen de mots dans une image). Si le score de similarité associé au lieu virtuel est le plus élevé, l'image courante a plus de mots en commun avec L_{-1} qu'avec n'importe quel autre lieu. Elle provient donc d'un lieu inconnu et un nouveau lieu est enregistré dans la base.

4.5 Prédiction

Une prédiction du lieu au temps k peut être obtenue à partir de la position antérieure et d'un modèle de mouvement simple. La caméra est portée par une personne donc nous ne récupérons pas de données odométriques. Nous faisons l'hypothèse que deux images successives dans la séquence vidéo proviennent de lieux adjacents. Si la caméra est dans le lieu L_i au temps $k-1$ il est probable qu'elle soit dans L_i, ou dans le voisinage de L_i au temps k. Il est aussi probable que l'observation courante provienne d'un nouveau lieu. Ces considérations sont prises en compte dans le calcul de la prédiction. Elles permettent de réduire les erreurs en assurant une cohérence temporelle des résultats.

4.6 Normalisation

La probabilité *a posteriori* est estimée en multipliant la vraisemblance et le terme de prédiction et en normalisant par :

$$p\left(Z_k|Z^{k-1}\right) = \sum_{i=-1}^{N_l} p\left(Z_k|L_i\right) p\left(L_i|Z^{k-1}\right) \tag{6}$$

N_l est le nombre de lieux déjà enregistrés dans la carte. On obtient ainsi la densité de probabilité sur l'ensemble des lieux. Le lieu reconnu correspond au maximum *a posteriori*.

5 Nouvelle formulation de la reconnaissance de lieu

L'algorithme de reconnaissance de lieu présenté dans la partie 4 classifie les images de la séquence en lieux discrets : les images sont triées suivant leur apparence globale. Nous pouvons utiliser ce résultat pour identifier les mots appartenant à de potentiels objets susceptibles de bouger. L'idée est d'utiliser plusieurs passages dans un même endroit et de détecter les différences géométriques entre ces passages. On suppose que ces différences résultent du déplacement d'un ou plusieurs objets entre deux passages de la caméra.

5.1 Intégration de la notion d'objet dans le modèle

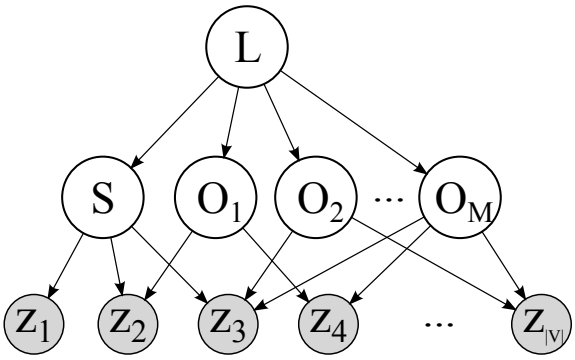


FIGURE 3 – Modèle graphique utilisant les notions d'objet et de structure statique dans la description de l'environnement

La modélisation des objets permet une mise à jour du formalisme de reconnaissance de lieu. Les mots visuels de l'observation courante appartiennent à la structure statique du lieu, ou à l'un des objets s'y trouvant. Le nouveau modèle graphique est illustré figure 3. Il met en évidence une nouvelle écriture de la probablité d'être dans le lieu L_i :

$$p\left(L_i|Z\right) = p\left(L_i\right) p\left(S|L_i\right) \prod_{l=1}^{M} p\left(O_l|L_i\right) \tag{7}$$

S est la structure statique du lieu. $\{O_l\}_{i=1}^{M}$ est l'ensemble des objets trouvés dans le lieu L_i. Les probabilités d'être dans la structure ou dans l'un des objets du lieu s'écrivent classiquement avec les probabilités d'occurence des mots visuels.

$$p\left(S|L_i\right) = \prod_{j=1}^{|V|} p\left(z_j|S, L_i\right) \tag{8}$$

$$p\left(O_l|L_i\right) = \prod_{j=1}^{|V|} p\left(z_j|O_l, L_i\right) \tag{9}$$

Le lieu est estimé au sens du maximum a posteriori.

5.2 Détection d'objets

La reconnaissance de lieu est exécutée en parallèle d'un algorithme de SLAM métrique estimant les coordonnées 3D des points de l'environnement et la trajectoire de la caméra [15]. Cet algorithme s'appuie sur l'utilisation d'images clefs que nous classons en lieux. Une fois l'image courante classifiée, la géométrie entre les points 3D du lieu et les points de l'image traitée est calculée. Nous utilisons les points reconstruits en 3D pour être plus robuste. Une procédure utilisant l'algorithme RANSAC calcule différentes transformations de caméras en appariant les points (mise en correspondance des descripteurs SURF) et choisit celle qui satisfait le maximum de points appariés. Cette étape permet dans un premier temps de vérifier le résultat de la reconnaissance de lieu. Si ces points ne vérifient pas de géométrie épipolaire le résultat est ignoré. Dans le cas contraire le résultat implique que les images ont la même structure et peut être utilisé dans un post-traitement. Les points vérifiant la géométrie sont supposés statiques, les autres sont les points des objets ayant bougés. On considère pour l'instant que la majorité des points sont statiques et que les points outliers proviennent d'un seul objet. La figure 5 montre deux exemples d'objets segmentés de cette manière.
Les points récupérés sont considérés appartenir à un même objet. On peut alors apprendre un modèle d'apparence de l'objet en sac de mot de la même façon qu'on apprend le modèle d'un lieu. Par la suite nous souhaitons détecter le déplacement de plusieurs objets et définir leur modèle d'apparence pour les identifier lors d'un second passage.

5.3 Résultats expérimentaux et discussion

Les figures 4 et 5 illustrent les résultats obtenus pour la reconnaissance de lieu sans la prise en compte des objets d'une part et la détection d'objets déplacés entre deux passages d'autre part. La figure 6 montre les résultats de l'algorithme FAB-MAP [7] sur l'une de nos séquences. La matrice donne la probabilité de provenir d'un même lieu : l'image 7 a une probabilité de 2% de provenir du même lieu que l'image 3 et une probabilité de 98% de provenir d'un nouveau lieu. Notre méthode vise à améliorer ce résultat en prenant en compte la notion d'objet.

FIGURE 4 – Reconnaissance de lieu : Exemple de paires d'images appartenant à un même lieu.

Nous avons pu lier les méthodes de SLAM métrique et topologique et identifier des objets. La prise en compte des objets dans la reconnaissance de lieu ne fournit pas

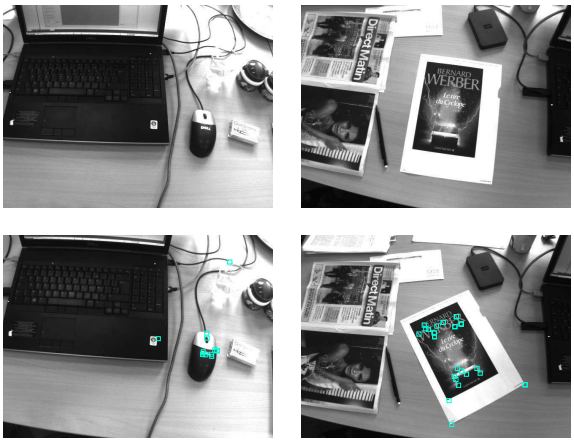


FIGURE 5 – Objets segmentés : le calcul de la géométrie met en évidence un ensemble de points (en bleu) ayant subi le même déplacement

FIGURE 6 – Résultats de l'algorithme FAB-MAP sur l'une de nos séquences. L'entrée (i, j) de la matrice est la probabilité que l'image i provienne du même lieu j. Les faibles valeurs sont en bleues, les fortes valeurs sont en rouges. Les fortes probabilités en dehors de la diagonale principale indiquent la reconnaissance d'un lieu déjà visité. Pour les valeurs de la diagonale, l'entrée (i, j) est la probabilité que l'image i provienne d'un nouveau lieu.

encore de résultats probants. Il est en particulier difficile d'identifier des points appartenant véritablement à un seul et même objet. La méthode utilisée ne fournit pour l'instant pas suffisamment de points pour créer un modèle pertinent, notamment lorsque l'objet subit une transformation affine importante. D'autre part, plusieurs cas sont à prendre en compte lors du déplacement d'un objet : l'objet est déplacé et reste dans le même lieu, l'objet disparait du lieu, l'objet apparait dans un lieu différent.

6 Conclusion

Après avoir présenté et expérimenté une première méthode de reconnaissance de lieu, nous avons mis en évidence l'intérêt d'associer des méthodes de SLAM métrique et topologique pour détecter le déplacement d'objets entre deux passages de la caméra. Les différentes explorations de la caméra dans un même environnement permettent de modéliser les objets détectés. Enfin nous avons proposé une nouvelle formulation caractérisant les lieux et les objets dynamiques d'un environnement. Nous souhaitons dans le cadre de la généralisation de ces travaux modéliser un environnement intérieur sujet à de nombreuses modifications : bâtiment, ateliers, grandes surfaces.

Références

[1] SungHwan Ahn, Minyong Choi, Jinwoo Choi, and Wan Kyun Chung. Data association using visual object recognition for EKF-SLAM in home environment. In *IROS*, pages 2588–2594, 2006.

[2] A. Angeli, D. Filliat, S. Doncieux, and J.-A. Meyer. A fast and incremental method for loop-closure detection using bags of visual words. *IEEE Transactions On Robotics, Special Issue on Visual SLAM*, 2008.

[3] Adrien Angeli and Andrew Davison. Live feature clustering in video using appearance and 3D geometry. In *BMVC*, pages 1–11, 2010.

[4] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. SURF : Speeded Up Robust Features. *9th European Conf on Computer Vision*, pages 404–417, 2006.

[5] Tom Botterill, Richard Green, and Steven Mills. A Bag-of-Words Speedometer for Single Camera SLAM. In *Image and Vision Computing New Zealand*, pages 1–6, Wellington, NZ, November 2009.

[6] R. O. Castle, G. Klein, and D. W. Murray. Combining monoslam with object recognition for scene augmentation using a wearable camera. 28(11) :1548 – 1556, 2010.

[7] Mark Cummins and Paul Newman. FAB-MAP : Probabilistic Localization and Mapping in the Space of Appearance. *The International Journal of Robotics Research*, 27(6) :647–665, 2008.

[8] Mark Cummins and Paul Newman. Highly scalable appearance-only SLAM : Fab-map 2.0. In *Robotics Science and Systems (RSS)*, Seattle, USA, June 2009.

[9] Andrew J Davison, Ian D Reid, Nicholas D Molton, and Olivier Stasse. MonoSLAM : real-time single camera SLAM. *IEEE transactions on pattern analysis and machine intelligence*, 29(6) :1052–67, June 2007.

[10] Vincent Gay-Bellile, Mohamed Tamaazousti, Romain Dupont, and Sylvie Naudet Collette. A vision-based hybrid system for real-time accurate localization in an indoor environment. In *Computer Vision Theory and Applications (VISAPP)*, pages 216–222, 2010.

[11] Kin Leong Ho and Paul Newman. Detecting loop closure with scene sequences. *International Journal of Computer Vision*, 74(3) :261–286, September 2007.

[12] Arnold Irschara, Christopher Zach, Jan-Michael Frahm, and Horst Bischof. From structure-from-motion point clouds to fast location recognition. In *CVPR*, pages 2599–2606. IEEE, 2009.

[13] Georg Klein and David Murray. Parallel tracking and mapping for small AR workspaces. In *Proc. Sixth IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR'07)*, Nara, Japan, November 2007.

[14] David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2) :91–110, November 2004.

[15] E. Mouragnon, M. Lhuillier, M. Dhome, F. Dekeyser, and P. Sayd. Monocular Vision Based SLAM for Mobile Robots. *18th International Conference on Pattern Recognition (ICPR'06)*, pages 1027–1031.

[16] D. Nister and H. Stewenius. Scalable Recognition with a Vocabulary Tree. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR'06)*, pages 2161–2168.

[17] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization : Improving particular object retrieval in large scale image databases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[18] Gerhard Reitmayr, Ethan Eade, and Tom Drummond. Semi-automatic annotations in unknown environments. In *Proc. ISMAR 2007*, pages 67–70, Nara, Japan, Nov. 13–16 2007.

[19] E. Royer, M. Lhuillier, M. Dhome, and T. Chateau. Localization in Urban Environments : Monocular Vision Compared to a Differential GPS Sensor. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pages 114–121.

[20] Grant Schindler, Matthew Brown, and Richard Szeliski. City-Scale Location Recognition. *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7, June 2007.

[21] Grant Schindler, Matthew Brown, and Richard Szeliski. City-Scale Location Recognition. *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7, June 2007.

[22] Josef Sivic and Andrew Zisserman. Video Google : a text retrieval approach to object matching in videos. *Proceedings Ninth IEEE International Conference on Computer Vision*, (Iccv) :1470–1477 vol.2, 2003.

[23] Hauke Strasdat, J. M. M. Montiel, and Andrew J. Davison. Real-time monocular slam : Why filter ? In *ICRA*, pages 2657–2664. IEEE, 2010.

[24] A. Torralba, K.P. Murphy, W.T. Freeman, and M.A. Rubin. Context-based vision system for place and object recognition. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 273–280 vol.1, 2003.

[25] Junqiu Wang, Hongbin Zha, and Roberto Cipolla. Coarse-to-Fine vision-based localization by indexing scale-invariant features. *IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics Society*, 36(2) :413–22, April 2006.